Provenance-Aware LOD Datasets for Detecting Network Inconsistencies

Leslie F. Sikos,[1] Dean Philp,[2] Shaun Voigt[2],
Catherine Howard,[2] Markus Stumptner,[1] Wolfgang Mayer[1]

[1] University of South Australia, Adelaide, Australia
[2] Defence Science and Technology Group, Adelaide, Australia

Abstract. Contextualized knowledge graphs (CKGs) have been gaining importance in recent years by providing context-aware datasets in various knowledge domains. In communication network analysis, for example, CKGs can be used to improve cyber-situational awareness or to reason about network topologies. Despite the potential of these graphs, there is a lack of published CKG-based datasets for communication networks. The complexity, scale, and rapid changes of real-world communication networks make it crucial to capture not only network knowledge in network datasets, but also additional metadata. Therefore, this paper presents communication network datasets, enriched with provenance, timestamps, and location data, which can be used for benchmarking, in silico experiments, and aimed at serving as the basis for further applications and research.

1 Introduction

Cyber-situational awareness applications rely on heterogeneous data sources, ranging from routing messages to router configuration files through to open datasets, all of which have different file formats and data structures [1]. The *Resource Description Framework (RDF)*[1] can be used to provide a uniform representation for network data derived from heterogeneous resources [2], however, automatically generated data may not be considered authoritative, verifiable, and reproducible, unless *data provenance* (the source or origin of data) is captured [3], optionally complemented by other types of metadata and the uncertainty and vagueness of statements about dynamic network knowledge [4]. Providing provenance for RDF statements is a long-standing, non-trivial problem in the Semantic Web research community, which led to different approaches. Some extended the standard RDF data model (e.g., *RDF*$^+$ [5], *SPOTL* [6], and *RDF** [7]) or the RDFS semantics (*Annotated RDF Schema* [8], *G-RDF* [9]), others proposed alternate data models (e.g., *N3Logic* [10]), decomposed RDF graphs (RDF molecule [11]), encapsulated provenance with RDF triples (e.g.,

[1] `https://www.w3.org/RDF/`

Provenance Context Entity (PaCE) [12], *singleton property* [13]), captured context (e.g., *named graphs* [14], *RDF triple coloring* [15], *nanopublications* [16]), and utilized vocabularies and ontologies, such as the *Provenir* ontology [17]and *NdFluents* [18]. While there are several ontologies described in the literature for network knowledge representation, very few, such as the *Situational Awareness (SAW) Ontology* [19] and the *Communication Network Topology and Forwarding Ontology (CNTFO)*[2] [20], are purposefully designed for capturing provenance-aware network knowledge for applications that require cyber-situational awareness. With the need for CKG-based communication network datasets in mind, as well as the lessons learned from popular datasets (e.g., DARPA '99 [21]), this paper presents novel CKG-based datasets. The presented datasets utilize named graphs to capture provenance, thereby differentiating between network knowledge statements (by source type), CNTFO terms to capture network knowledge and network-specific provenance, and *PROV-O*[3] to describe general provenance.

2 Provenance-Aware Network Knowledge Datasets

Using the publicly available *Common Open Research Emulator (CORE)*[4], realistic scenarios are modeled in these datasets, in which two Australian businesses—each with sites in Adelaide and in Melbourne—require two Internet Service Providers (ISPs) and 24/7 Internet access (*dual-homing*). The underlying model consists of 60 devices in total, each with several network interfaces. Two types of network models have been constructed (8 models in total), covering IPv4 and IPv6 base cases and well-documented deliberate misconfigurations, the latter of which are errors that impact both network performance and security. These models were used to generate context-aware RDF datasets, collectively called *ISPnet*. These datasets are compliant with Semantic Web best practices and constitute LOD data. All nodes of the corresponding RDF graphs are globally dereferencable. The integrity of the datasets have been checked with HermiT,[5] FaCT++,[6] and Pellet.[7] This paper focuses on two of these publicly released datasets: 1) IPv4 base[8] and 2) IPv4 overlapping subnets.[9] They cover heterogeneous network data derived from device configurations, traceroutes, OSPF LSAs, and arpings. The DL expressivity of both datasets is $\mathcal{ALU}$. The first dataset defines 55 classes and 322 individuals with 1,595 axioms. The second dataset has 14 classes and 295 individuals defined in the form of 1,264 axioms. The dataset files are accompanied by standard-compliant VoID[10] descriptions.

[2] `http://purl.org/ontology/network/`
[3] `http://www.w3.org/ns/prov-o`
[4] `https://www.nrl.navy.mil/itd/ncs/products/core`
[5] `http://www.hermit-reasoner.com`
[6] `http://owl.cs.manchester.ac.uk/tools/fact/`
[7] `https://github.com/stardog-union/pellet`
[8] `http://purl.org/dataset/ispnet/base/`
[9] `http://purl.org/dataset/ispnet/overlap/`
[10] `https://www.w3.org/TR/void/`

3 Case Study

We provide an excerpt from our two datasets, namely, ISPnet and ISPnetOL. The ISPnet dataset was generated using our base network model, whereas ISPnetOL was generated using a deliberate misconfiguration of the base network model. Both datasets contain four types of named graphs that correspond to heterogeneous network data sources (CORE, traceroute, arping, and OSPF LSAs). The datasets demonstrate three levels of provenance: triple-level, graph-level, and dataset-level provenance. Triple-level provenance includes statements such as `ispnet:C1-ADL-R1 prov:atLocation dbpedia:Adelaide`, indicating that Router 1 of Customer 1 is geographically located in Adelaide. Graph-level provenance includes statements such as `ispnet:TRACEROUTE4 net:ImportHost "C1-ADL-PC3"`, which indicates that Computer 3 of Customer 1 is where the traceroute command was executed. Dataset-level provenance includes statements such `<http://purl.org/dataset/ispnet/base/> prov:wasAssociatedWith "DST Group Australia".`

By comparing the CORE graphs between the two datasets, it can be inferred that C1-ADL-R1_eth1 was connected to 10.10.0.164/30 on 13 May 2018, whereas on 14 May the connection changed to 10.10.0.185/29; this is the first indication of a configuration error (see Fig. 1).

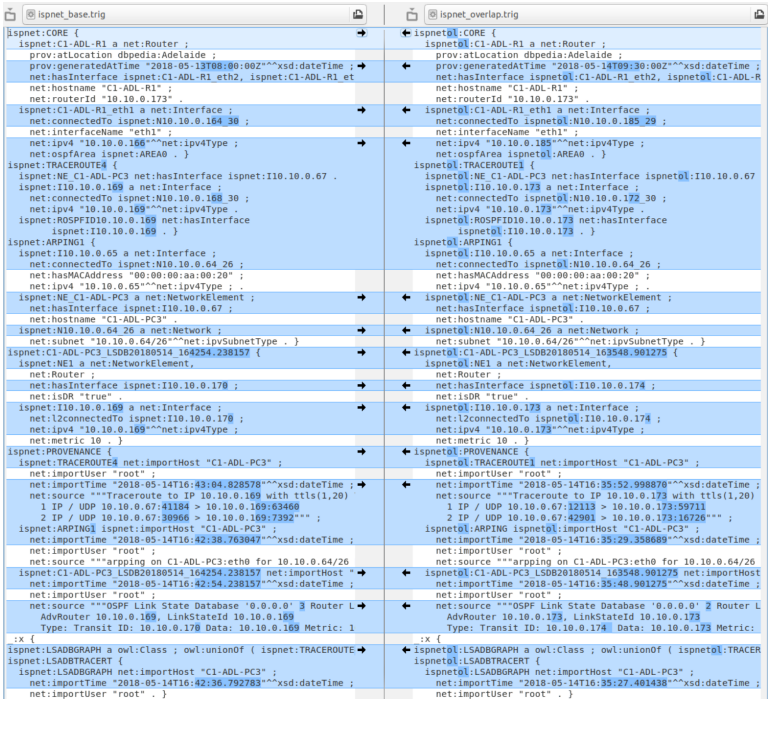

Fig. 1. Comparison of the base case and the misconfiguration using provenance

By comparing the PROVENANCE graphs in conjunction with the TRACEROUTE, ARPING and LSDB graphs, it can be inferred that C1-ADL-PC3 could previously reach 10.10.0.169 (the customer gateway), but subsequently could access only 10.10.0.173 (not the gateway).

This allows another inference, namely, that Customer 1 in Adelaide has lost Internet access, which is important for cyber-situational awareness. Without statements of three facets of provenance, i.e., time, location, and `importHost`, we could not have performed the required information fusion and reasoning to make this inference. Importantly, this inference is indeed correct: our specific deliberate misconfiguration example actually does cause Customer 1 to lose Internet access.

Figure 2 shows a small part of the RDF graph of the first dataset file of the case study, demonstrating statements derived from three different data sources (CORE, a traceroute, and an arping), and some of the associated provenance statements.

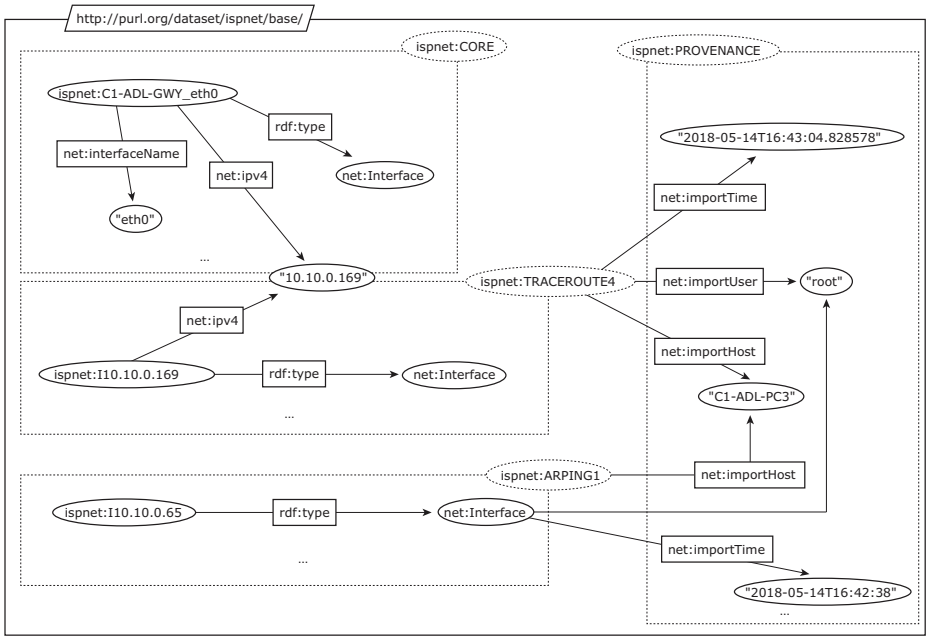

Fig. 2. Part of the RDF graph of the case study representing network knowledge graphs and a provenance graph

The statements about the IP address associated with interface `C1-ADL-GWY_eth0` and `I10.10.0.169` suggest that these entities are actually identical (a link can be created between the two using `owl:sameAs`), only they were named differently at different stages of network knowledge discovery based on the information available at the time. The automated identification of

such relationships is beneficial for network analysts and enables the generation of useful, non-trivial RDF statements that help understand network element connectivity and traffic flow.

4 Conclusion

Due to the unavailability of CKG-based datasets for communication networks, practitioners and researchers need standard datasets to compare, contrast, and build upon to further both practical applications and research. This paper presented such context-aware network knowledge datasets, which can be used for modeling communication networks and testing semantic formalisms for capturing metadata-enriched network knowledge statements with RDF quadruples. These datasets are novel in terms of complexity, statement-level and statement group-level metadata, realistic environment model, and configuration parameters. They cover heterogeneous network data derived from a variety of sources, which can be utilized for facilitating information fusion.

References

1. Sikos, L. F. (ed.) (2018). AI in Cybersecurity. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-98842-9
2. Sikos, L. F., Stumptner, M., Mayer, W., Howard, C., Voigt, S., Philp, D. Summarizing Network Information for Cyber-Situational Awareness via Cyber-Knowledge Integration. AOC 2018 Convention, Adelaide, Australia, May 2018
3. Sikos, L. F., Stumptner, M., Mayer, W., Howard, C., Voigt, S., Philp, D. (2018) Automated reasoning over provenance-aware communication network knowledge in support of cyber-situational awareness. In: Liu, W., Giunchiglia, F., Yang, B. (eds.) Knowledge Science, Engineering, and Management. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-99247-1_12
4. Sikos, L. F. Handling Uncertainty and Vagueness in Network Knowledge Representation for Cyberthreat Intelligence. 2018 IEEE World Congress on Computational Intelligence, Rio de Janeiro, Brazil, July 2018
5. Dividino, R., Sizov, S., Staab, S., Schueler, B. (2009) Querying for provenance, trust, uncertainty and other meta knowledge in RDF. Web Semant. Sci. Serv. Agents World Wide Web 7(3):204–219. https://doi.org/10.1016/j.websem.2009.07.004
6. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G. (2012) YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell. 194:28–61. https://doi.org/10.1016/j.artint.2012.06.001
7. Hartig, O., Thompson, B. (2014) Foundations of an alternative approach to reification in RDF. https://arxiv.org/abs/1406.3399
8. Zimmermann, A., Lopes, N., Polleres, A., Straccia, U. (2012) A general framework for representing, reasoning and querying with annotated Semantic Web data. Web Semant. Sci. Serv. Agents World Wide Web 11:72–95. https://doi.org/10.1016/j.websem.2011.08.006
9. Analyti, A., Damsio, C.V., Antoniou, G., Pachoulakis, I. (2014) Why-provenance information for RDF, rules, and negation. Ann. Math. Artif. Intell. 70(3):221–277. https://doi.org/10.1007/s10472-013-9396-0

10. Berners-Lee, T. (2008) Notation 3 Logic. https://www.w3.org/DesignIssues/N3Logic. Accessed 3 April 2018
11. Ding, L., Finin, T., Peng, Y., Da Silva, P.P., McGuinness, D.L. (2005) Tracking RDF graph provenance using RDF molecules. Fourth International Semantic Web Conference, Galway, Ireland, 6–10 November 2015
12. Sahoo, S.S., Bodenreider, O., Hitzler, P., Sheth, A., Thirunarayan, K. (2010) Provenance Context Entity (PaCE): scalable provenance tracking for scientific RDF data. In: Gertz, M., Ludscher, B. (eds.) Scientific and statistical database management. Lect. Notes Comput. Sci., vol. 6187, pp. 461–470. Heidelberg: Springer. https://doi.org/10.1007/978-3-642-13818-8_32
13. Nguyen, V., Bodenreider, O., Sheth, A. (2014) Don't like RDF reification? Making statements about statements using singleton property. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 759–770. New York: ACM. https://doi.org/10.1145/2566486.2567973
14. Carroll, J.J., Bizer, C., Hayes, P., Stickler, P. (2005) Named graphs, provenance and trust. In: Proceedings of the 14th International Conference on World Wide Web, pp. 613–622. New York: ACM. https://doi.org/10.1145/1060745.1060835
15. Flouris, G., Fundulaki, I., Pediaditis, P., Theoharis, Y., Christophides, V. (2009) Coloring RDF triples to capture provenance. In: Bernstein, A., Karger, D. R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) The Semantic Web – ISWC 2009. Lect. Notes Comput. Sci., vol. 5823, pp. 196–212. Heidelberg: Springer. https://doi.org/10.1007/978-3-642-04930-9_13
16. Groth, P., Gibson, A., Velterop, J. (2010) The anatomy of a nanopublication. Inform. Serv. Use 30(1–2):51–56. https://doi.org/10.3233/ISU-2010-0613
17. Sahoo, S.S., Sheth, A. (2009) Provenir ontology: towards a framework for eScience provenance management. Microsoft eScience Workshop, Pittsburgh, PA, USA, 15–17 October 2009
18. Gimnez-Garca, J.M., Zimmermann, A., Maret, P. (2017) NdFluents: an ontology for annotated statements with inference preservation. In: Blomqvist, E., Maynard, D., Gangemi, A., Hoekstra, R., Hitzler, P., Hartig, O. (eds.) The Semantic Web. Lect. Notes Comput. Sc., vol. 10249, pp. 638–654. Cham: Springer. https://doi.org/10.1007/978-3-319-58068-5_39
19. Sheth, A. (2007) Leveraging Semantic Web techniques to gain situational awareness. Can Semantic Web techniques empower perception and comprehension in cyber situational awareness? Cyber Situational Awareness Workshop, Fairfax, VA, USA, 14-15 Nov 2007.
20. Sikos, L. F., Stumptner, M., Mayer, W., Howard, C., Voigt, S., Philp, D. (2018) Representing Network Knowledge Using Provenance-Aware Formalisms for Cyber-Situational Awareness. Procedia Computer Science
21. Thomas, C., Sharma,V., Balakrishnan, N. (2008) Usefulness of DARPA dataset for intrusion detection system evaluation. In: Proceedings of the 2008 SPIE Defense and Security Symposium, Orlando, FL, USA, 17 March 2008. https://doi.org/10.1117/12.777341